INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 4 (2016) PAGES 1150 - 1156
RECEIVED : 19/08/2016. PUBLISHED : 27/08/2016

August 27, 2016

SECURE AND AUTHORIZED DEDUPLICATION DATA IN CLOUD STORAGE SYSTEMS

[1]Sakthivelu. M , [2]Geetha. S,
[1]M.Phil Scholar, Dept of Computer Science, Muthurangam Arts College, Vellore,
[2]Asst.Prof, Dept of Computer Science, Muthurangam Arts College, Vellore.

Abstract:
Now days cloud storage is one of most important aspects that manage space and provide secure data in hybrid cloud systems. The Deduplication is the technique that provide the data in the compression form and that eliminate the duplication or repeating files in the cloud storage. Some time no of users become access the cloud may possible restore the same file again and again into cloud storage area that increase the storage space that give headache to provider to manage space. By storing no of unique copy of files cloud providers face the problem of data transfer cost. The existing Deduplication system becomes high cost in terms of new security is lesser. In this research proposed new Deduplication method to secure cloud data and avoid the data redundancy in the cloud storage. In this approach provides efficient space in cloud with chunk – level Deduplication and data privacy. This approach proposed new convergent key encryption to avoid data repeating groups. The convergent encryption method has been used to encrypt the user's data earlier than uploading, to protect the privacy of the data. Different from the conventional Deduplication system, the dissimilar rights of users are additional well thought-out in duplicate check. This research will shows reduce overhead of cloud provider storage and computational cost.

Keywords – Deduplication, Cloud, computational cost.

1. INTRODUCTION

In recent trends cloud systems are potentially storage space provide by the cloud providers and cloud users can use much space they can seller the entire time look for minimized storage and cost in the cloud system. A system which has been extensively accepted is user Deduplication. The straightforward idea in the rear Deduplication is to stock up replacement data (either files or blocks) only once. Consequently, if a consumer needs to upload a file (block) which is previously stored, the cloud provider must add with user and owner list into the block. Deduplication has confirmed to accomplish elevated space and cost savings and many cloud storage providers are at presently approve it. Deduplication system can used avoid 90 to 95% unnecessary storage space in the cloud system.

1.1 Levels of Deduplication

- File Level Deduplication
- Block Level Deduplication

1.1.1 File Level Deduplication

File-level Deduplication watches for multiple copies of the same file, stores the first copy, and then just links the other references to the first file. Only one copy gets stored on the disk/tape archive. Ultimately, the space you save on disk relates to how many copies of the file there were in the file system.

1.1.2 Block Level Deduplication

Block-level, sometimes are called variable block-level Deduplication, looks at the data block itself to see if another copy of this block already exists. If so, the second (and subsequent) copies are not stored on the disk/tape, but a link/pointer is created to point to the original copy. To make de-duplication secure to apply certain security mechanism like encryption. Traditional encryption requires different users to encrypt their data with their own keys, so identical data copies of different users will lead to different cipher text and for this reason de-duplication is incompatible with traditional encryption. Convergent encryption provides a possible option to implement data confidentiality while realizing Deduplication. The Main motivation of this research is Deduplication in the Cloud Storage systems because.*Cost***:** Add up the cost of what your system will need in order to enable Deduplication. Cloud user need provide much higher cost for cloud space.*Tape:* If you are migrating all of your data to tape on a daily basis, then the need for Deduplication is moot. *Data footprint***:** If you have a small data footprint, less than 10TBs, the need for Deduplication can be balanced against the cost. With disk costs decreasing, a single drawer in your storage array may be able to handle the load without Deduplication.

2. LITERATURE REVIEW

Literature review is the process of presenting the summary of the journal articles, conference papers and study resources. So in this section have studied the related topics a summarized it below.Mark W. Storer Kevin Greenan Darrell et al [1], proposed new scheme secure deduplication models: authenticated and anonymous. The security property most associated with encryption is *secrecy*, which states that only authorized users are able to read plaintext data.This method is based on Context of the Farsite system.The main drawback of the system is that both well-behaved and malicious users are anonymous. Danny Harnik et all[2], presented new approach cross-user source based-deduplication. That provide clients of cloud storage services can do so, regardless of the service they use, by encrypting their data before the local client software of the storage service operates on their files.This method is based on Cross – User Deduplication.The main advantages of the system is Reducing the risk of data Leak.The main drawback of system is that it eliminates all bandwidth savings of deduplication, and that the service provider and/or the users must pay for transferring the raw amount of data. Shai Halevi et al[3], introduced new method name Prof of Ownership (PoWs). A proof mechanism that notes that this is somewhat similar to proofs of irretrievability (PORs) and proofs of data possession (PDPs) with a role reversal (the client is the prover rather than the server).This system is based on method of Prof of Ownership Protocol (PoWs).The main advantage of the system This Method is suitable for the attack scenarios of common hash functions, malicious software, or accidental leakage.Drwaback of the system is When the files grows this is not better deduplication ratio as well as for slower networks. Iuon-Chang Lin presents a new approach deduplication in cloud storage. These approach categories of data deduplication strategy, and extend the fault-tolerant digital signature scheme proposed by Zhang on examining redundancy of blocks to achieve the data deduplication. This method is based on the system of Zhang’s scheme.The main advantage of the system improves the speed of data deduplication phase by calculating a feature value to represent every column and row instead of verifying the hash value of every block. Traditional backup contains data streams with locality of reference.The system is based on the method of Enhanced Dynamic Whole File De-Duplication (DWFD).

Iuon-Chang Lin et al[4], presents a new approach deduplication in cloud storage. These approach categories of data deduplication strategy, and extend the fault-tolerant digital signature scheme proposed by Zhang on examining redundancy of blocks to achieve the data deduplication.This system is based on the method of Zhang’s scheme.The main advantage of the system is the speed of data deduplication phase by calculating a feature value to represent every column and row instead of verifying the hash value of every block.The main drawback of the system is worst case in that cloud storage server will regard all blocks as a new blocks and store all of these blocks. **:** Jia Xu article presents new method to cloud storage to save bandwidth and storage.The system is based on the method of Convergent Cryptosystem. Jan Stanek etal[5], present new multi layered Cryptosystem. All files are initially declared unpopular and are encrypted with two layers the inner layer is applied using a convergent cryptosystem, whereas the outer layer is applied using a semantically secure threshold cryptosystem. This system is based on the method of Convergent Threshold Cryptosystem. **:** M. Shyamala Devi et al[6], present about the backup of the private storage cloud belongs to the non-traditional backup.This system is based on the method of Enhanced Dynamic Whole File De-Duplication (DWFD).The main advantage of the system To optimize the private cloud storage backup in order to provide high throughput to the users of the organization by increasing the de-duplication efficiency.

2.1 Comparison of Various Data Deduplication Approaches

Deduplication Approach	**Bandwidth Utilization**	**Storage Utilization**	**Throughput**	**Deduplication Ratio**	**Efficiency**	**Cost**
File Level	*Low*	*Medium*	*High*	*Low*	*Average*	*Low*
Block Level	*Medium*	*High*	*Low*	*High*	*High*	*Medium*
Source Based	*Low*	*Medium*	*Medium*	*Medium*	*Medium*	*Low*
Target Based	*High*	*High*	*Medium*	*Medium*	*Medium*	*|High*
Inline	*Low*	*Low*	*Low*	*Low*	*Medium*	*Low*
Post Process	*High*	*Low*	*Medium*	*High*	*High*	*High*

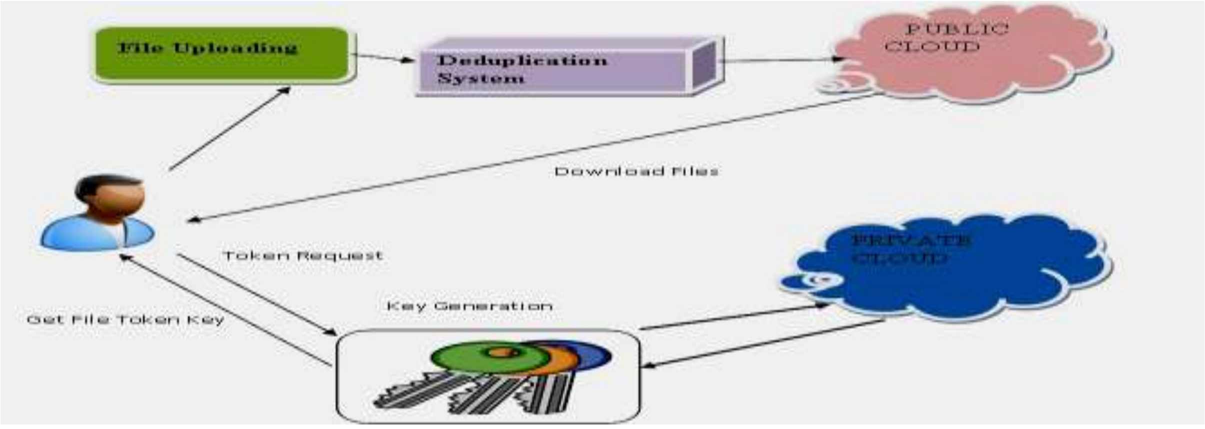
Table.1.Comparison of Various Data Deduplication Approaches.

3. PROPOSED APPROACH

The hybrid cloud is a combination of the both public cloud and private cloud. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. The users cannot share these private keys of rights, which mean that it can prevent the privilege key sharing among users in the above straightforward construction. The user needs to send a request to the private cloud server to get a file token. To perform the duplicate check for some file, the user needs to get the file token from the private cloud server. The private cloud server will check the user's identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading the file. The user either uploads the file or runs PoW based on the results of duplicate check. If this approach is used the public cloud this method can't provide the security for our private data and hence our private data will be loss.

3.1 Secure Deduplication System

There are three entities defined in this system, User, Private cloud and Secure cloud service provider(S-CSP) in public cloud.



3.2 Proposed Deduplication Model

Let D a system that finds out duplicate copies of the file using Authorized deduplication system in hybrid cloud.

D = {F, B, C, FT, CK, M, O} Where,

F – Is the file {f1, f2, f3,…fn}
F1 – is the Block {B1, B2, B3…Bn}
B1 – {CB$_i$, FTBi, CKi} where CBi- is Set of cipher text Block.
FT- File token, CK – Convergent Key, M- Metadata of File
O – Output Consists of reduce Database size.

3.2.1. Step by Step Proposed Model

Step 1: *File F is divided into multiple blocks F= ΣBi, F = size (F) / 4096.*

Step 2: *KeyGen (1 λ) →k is convergent key generation algorithm, generate secrete key using security parameter 1 λ .Secret key stores in internal DB of Security Service (SS).*

Step 3: *Enc (k,F)→C is encryption algorithm that takes secrete key k and then file and then F output is cipher text C.*

Step 4:*Generate File Token FT for each block.*

Step 5:*Dec (k,C)→F is Decryption algorithm that takes secrete key k and ciphertext C and then output is original file F. F=ΣPlainText (Bi).*

Step 6*: Detect duplication. Security Service generates TiBi file Token on basic on Bi, If the same Bi comes in then it will generate the same TiBi. i.e. TiBi = File token generation (Bi); Then it will store the TiBi to the Own Security Db. If file is found in database it generates response.*

3.3 Proposed System Advantages

- The encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same cipher text.
- To prevent unauthorized access, a secure proof of ownership (POW) protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found.
- Data confidentiality is maintained.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section to evaluate the overhead introduced by our system in terms of storage space and computational complexity. Also evaluate proposed method resilience against potential attacks. In order to refer to a real scenario, here use the same parameters of Dutch T Meyer et al are used, but our calculations hold true for other scenarios.

4.1 Private Cloud

The private cloud to provide the user access and generate the file token who is asking to upload and download file from the public cloud.

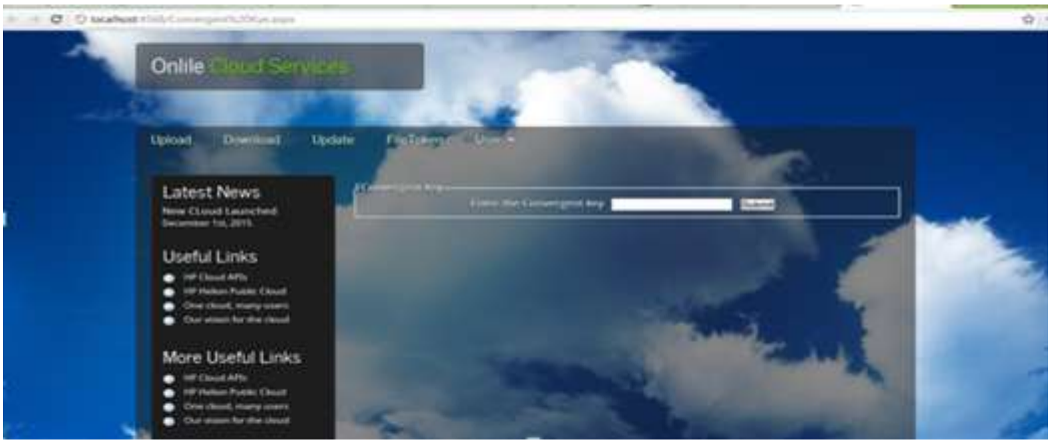


Fig.2. Convergent Key Generation

4.1.1 Storage

The first step of the storage protocol requires the server to encrypt Bi, Ki and Si. As the encryption is symmetric, the cost of each encryption can be considered constant, so for N blocks the total cost is O(N). The second step of the protocol requires the metadata manager to hash each block in order to compare it with the ones already stored. The cost of this operation is O(logM) and it is performed for each block. The cost of the update of the data structures can be considered constant.

4.1.2 Retrieval

The first step of the retrieval protocol requires the metadata manager to compute a hash of the concatenation of user id and file name. The cost of this operation can be considered constant. Even the lookup in the file table, in order to get the pointer to the first block of the file, has a constant cost

4.1.3Deduplication Rate

Our proposed solution aims to provide a robust security layer which provides confidentiality and privacy without impacting the underlying deduplication technique.

Using the gathered information so far, able to estimate the cost of backing up all the machines have to Amazon S3 over a one year period. Another example shows that the storage space for the cloud service providers is compared with the existing storage with the deduplication storage. More storage space is reduced in the deduplication scheme.

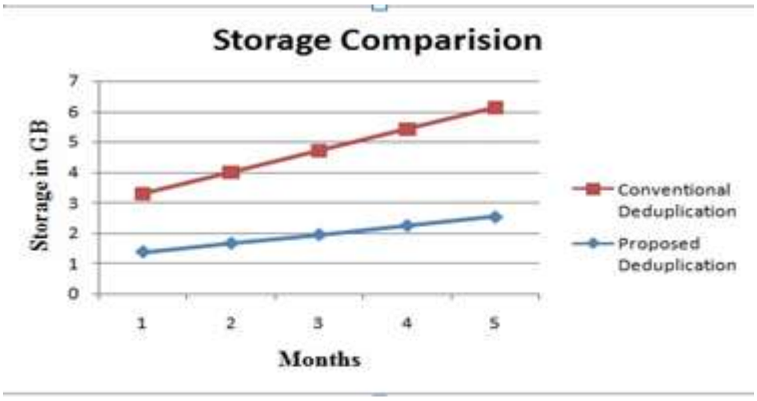


Fig.3. File Storage in Cloud Comparison

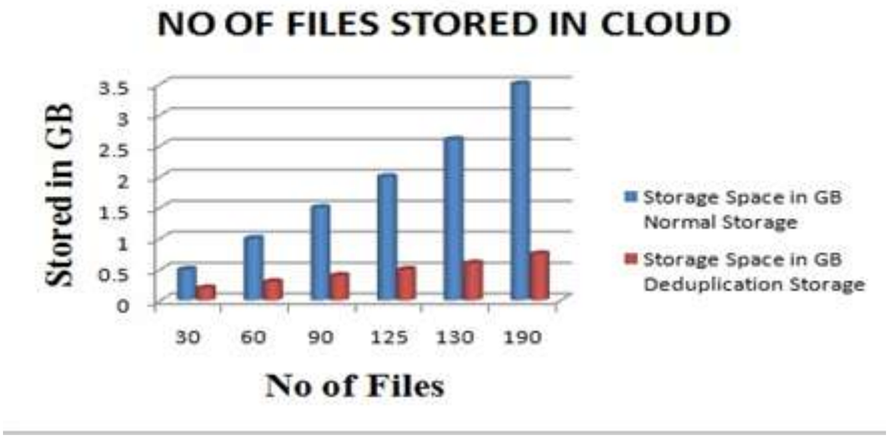


Fig.4. No of Files Stored Comparison

5. CONCLUSION AND FUTURE WORK

Secure and authorized Deduplication system in cloud which accomplishes discretion and allows block-level deduplication at the same time. In this system is built on top of convergent encryption. The Results that it is worth performingfilelevel deduplicationinstead of block-level deduplication since the gains in terms of storage space are not affected by the overhead of metadata management, which is minimal. Additional layers of encryption are added by the server and the optional HSM. As the additional encryption is symmetric, the impact on performance is negligible. This system showed that our design, in which no component is completely trusted, prevents any single component from compromising the security of the whole system. Our solution also prevents curious cloud storage providers from inferring the original content of stored data by observing access patterns or accessing metadata. Furthermore, this approach showed that our solution can be easily implemented with existing and widespread technologies. Finally, this solution is fully compatible with standard storage APIs and transparent for the cloud storage provider, which does not have to be aware of the running deduplication system. Therefore, any potentially untrusted cloud storage provider such as Amazon, Dropbox and Google Drive, can play the role of storage provider.In future this work can extend more security such as proof of possession, Data Integrity checking and search over Encrypted

Data in the cloud storage. This approach mainly focused about storage and retrieval data in cloud but in the future these will extents with edit and delete the files in the cloud storage.

REFERENCES

[1]. Mark W Storer, Kevin Greenan, Darrell DE Long, and Ethan L Miller. "*Secure data deduplication*". In Proceedings of the 4th ACM international workshop on Storage security and survivability, pages 1–10. ACM, 2008.

[2]. Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg."*Side channels in cloud services: Deduplication in cloud storage. Security & Privacy"*, IEEE, 8(6):40–47, 2010.

[3]. Shai Halevi, Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg," *Proofs of Ownership in Remote Storage Systems*", Proceedings of the 18th ACM conference on Computer and communications securityPages491-500-2011.

[4]. Iuon-Chang Lin and Po-Ching Chien," *Data Deduplication Scheme for Cloud Storage",International Journal of Computer, Consumer and Control (IJ3C), Vol. 1, No.2 (2012).*

[5]. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "*A secure data deduplication scheme for cloud storage*," Tech. Rep. IBM Research, Zurich, ZUR 1308-022, 2013.

[6].M. Shyamala Devi, V. Vimal Khanna, and A. Naveen Bhalaji," *Enhanced Dynamic Whole File De-Duplication (DWFD) for Space Optimization in Private Cloud Storage Backup"* International Journal of Machine Learning and Computing, Vol. 4, No. 4, August 2014.